# Estimation d'un modèle à blocs latents par l'algorithme SEM

Christine Keribin[1,3] & Gérard Govaert [2] & Gilles Celeux [3]

[1] *Laboratoire de Mathématiques UMR* 8628*, Université Paris-Sud* 11*, F*-91405 *Orsay cedex*
[2] *UMR* 6599*, CNRS et Université de Technologie de Compiègne, F*-60205 *Compiègne*
[3] *INRIA Saclay Île de France Projet* SELECT*, Bat* 425*, Université Paris-Sud* 11*, F*-91405 *Orsay cedex*

RÉSUMÉ

Les modèles de mélanges peuvent être utilisés pour résoudre le problème de la classification non supervisée simultanée d'un ensemble d'objets et d'un ensemble de variables. Le modèle à blocs latents définit une loi pour chaque croisement de classe d'objets et de classe de variables, et les observations sont supposées indépendantes conditionnellement au choix des classes d'objets et de variables. Mais il n'est pas possible de factoriser la loi jointe conditionnelle des labels et l'étape d'estimation de l'algorithme EM n'est pas calculable directement. Govaert et Nadif (2008) en ont proposé une approximation variationnelle qu'ils ont confrontée à un algorithme CEM. Nous présentons ici, dans le cadre de données binaires, l'utilisation d'un algorithme SEM effectuant l'étape d'estimation par échantillonneur de Gibbs, et nous comparons les résultats avec ceux des méthodes précédentes.

SUMMARY

Mixture models can be used to deal with the simultaneous clustering problem on a set of objects and a set of variables. The latent block model defines a distribution for each combinaison of an object label and a variable label and the data are supposed to be independent, given the object labels and the variable labels. But the factorization of the joint distribution of the labels, conditionally to the observed data, is not tractable, and the E-step of the EM algorithm cannot be performed. Govaert et Nadif (2008) proposed a variational approximation, which they compared to a CEM algorithm. We present here, in the case of binary data, a SEM algorithm, using a Gibbs sampling for the estimation step, and we compare the results with the other methods.

MOTS-CLES : Apprentissage et classification – Analyse de données et data mining
EM stochastique - Échantillonnage de Gibbs. Approximation variationnelle en champ moyen.

# 1 Introduction

Soit $\mathbf{x} = \{(x_{ij}); i \in I, j \in J\}$ une matrice de données de dimension $n \times d$ , où $I$ est un ensemble de $n$ objets (observations, cas) et $J$ un ensemble de $d$ variables (colonnes, attributs). Le but est d'opérer des permutations sur les objets et les variables pour construire

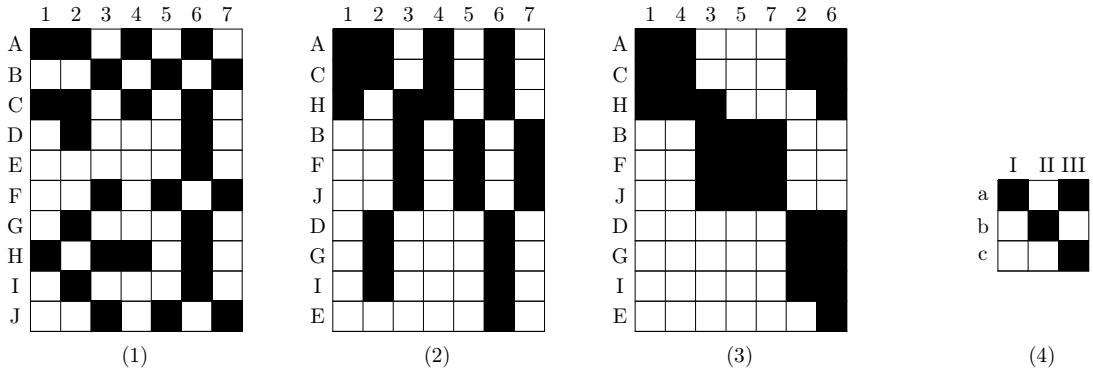


FIG. 1 – *Matrice de données binaires (1), réorganisées en partition sur $I$ (2), en partitions simultanées sur $I$ et $J$ (3), et résumé des données binaires (4)*

une structure de correspondance sur $I \times J$.

L'un des avantages des méthodes de classification par blocs est qu'elles permettent de réduire la matrice de données initiale $\mathbf{x}$ en une matrice plus simple ayant la même structure. Dans l'exemple représenté en Fig. 1 et proposé par Govaert et Nadif (2008), la matrice de taille $(n \times d) = (10 \times 7)$ est réduite en une matrice de taille $(g \times m) = (3 \times 3)$, où chaque valeur correspond à 1 ou 0. De plus, ces méthodes sont nettement plus rapides que des méthodes qui traitent les deux ensembles de façon séparée. Govaert et Nadif (2008) font une revue des procédures de détection des motifs dans des matrices de données, et comparent deux méthodes utilisant les modèles de mélange, dont un algorithme EM modifié (appelé BEM). Nous proposons ici une alternative à l'algorithme BEM, que nous appelerons SEM-Gibbs, et nous les comparons.

La partition $\mathbf{z}$ d'un échantillon $I$ en $g$ classes sera représentée par la matrice de classification $(z_{ik}, i = 1, \ldots, n, k = 1, \ldots g)$ où $z_{ik} = 1$ si $i$ appartient à la classe $k$ et 0 sinon. De façon similaire, la partition $\mathbf{w}$ d'un échantillon $J$ en $m$ classes sera représentée par la matrice de classification $(w_{j\ell}, j = d, \ldots, n, \ell = 1, \ldots m)$ où $w_{j\ell} = 1$ si $j$ appartient à la classe $\ell$ et 0 sinon. Les variables aléatoires seront notées en majuscule.

# 2 Modèle à blocs latents

Chaque coefficient $x_{ij}$ de la matrice $\mathbf{x}$ est le résultat du tirage d'une variable aléatoire $X_{ij}$. Dès que $\mathbf{z}$ et $\mathbf{w}$ sont fixés, la densité de la variable $X_{ij}$ est $\varphi(.; \alpha_{k\ell})$. Comme dans l'analyse en classes latentes, nous supposons l'indépendance conditionnelle des $n \times d$ variables $X_{ij}$ :

$$f(\mathbf{x}|\mathbf{z},\mathbf{w},\theta) = \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}.$$

Le modèle à blocs latents peut être défini comme un modèle de mélange

$$f(\mathbf{x},\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z};\theta)p(\mathbf{w};\theta)f(\mathbf{x}|\mathbf{z},\mathbf{w};\theta)$$

où $\mathcal{Z}$ et $\mathcal{W}$ représentent les ensembles de toutes les assignations possibles $\mathbf{z}$ de $I$ et $\mathbf{w}$ de $J$.

Dans le cas des données binaires, définissons le paramètre $\theta = (\pi, \rho, \alpha_{k\ell}; k = 1,\dots g, \ell = 1,\dots,m)$, où $\pi = (\pi_1,\dots,\pi_g)$ et $\rho = (\rho_1,\dots,\rho_m)$, pour obtenir le modèle à blocs latents de Bernoulli

$$f(\mathbf{x},\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

où $\alpha_{k\ell} \in (0,1)$ et $\varphi(x_{ij};\alpha_{k\ell}) = (\alpha_{k\ell})^{x_{ij}}(1-\alpha_{k\ell})^{1-x_{ij}}$.

# 3 L'algorithme BEM

L'algorithme EM (Dempster (1977)) maximise itérativement $Q(\theta,\theta^{(c)})$, l'espérance conditionnelle de la log-vraisemblance complète $L_C(\mathbf{z},\mathbf{w},\theta)$, en $\theta$, étant donné une estimation précédente de $\theta^{(c)}$ et les données observées $\mathbf{x}$ :

$$Q(\theta,\theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log \pi_k + \sum_{j,\ell} t_{j\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} e_{i,j,k,\ell}^{(c)} \log \varphi(x_{ij};\alpha_{k\ell})$$

où

$$s_{ik}^{(c)} = P(Z_{ik} = 1|\theta^{(c)},\mathbf{X} = \mathbf{x}), \quad t_{j\ell}^{(c)} = P(W_{j\ell} = 1|\theta^{(c)},\mathbf{X} = \mathbf{x})$$

et

$$e_{i,j,k,\ell}^{(c)} = P(Z_{ik} = 1, W_{j\ell} = 1|\theta^{(c)},\mathbf{X} = \mathbf{x}).$$

À cause de la structure de dépendance des variables $X_{ij}$, les valeurs $e_{i,j,k,\ell}^{(c)}$ ne peuvent pas être calculées analytiquement. Govaert and Nadif (2008) ont utilisé l'interprétation variationnelle de l'algorithme EM pour approximer l'étape E d'estimation. En effet, calculer la loi $p(\mathbf{w},\mathbf{z}|\mathbf{x};\theta^{(c)})$ dans l'étape E revient à maximiser l'énergie libre $\mathcal{F}(q_{wz})$ en $q_{wz}$

$$\mathcal{F}(q_{wz}) = \mathbb{E}_{q_{wz}}\left(\log \frac{p(\mathbf{x},\mathbf{w},\mathbf{z};\theta^{(c)})}{q_{wz}(\mathbf{w},\mathbf{z})}|\mathbf{x}\right).$$

Au maximum, $q_{wz}(\mathbf{w},\mathbf{z}) = p(\mathbf{w},\mathbf{z}|\mathbf{x};\theta^{(c)})$, si la fonction $q_{wz}$ est recherchée parmi l'ensemble des lois possibles. Quand la structure de covariance en $\mathbf{w}$ et $\mathbf{z}$ est trop compliquée, on peut

rechercher une approximation de $p(\mathbf{w}, \mathbf{z}|\mathbf{x}; \theta^{(c)})$ parmi les lois $q_{wz}(\mathbf{w}, \mathbf{z}) = q_w(\mathbf{w})q_z(\mathbf{z})$ se factorisant en $\mathbf{w}$ et $\mathbf{z}$ : c'est l'approximation en champ moyen (voir Keribin (2009) pour une revue de l'utilisation des méthodes variationnelles).
L'avantage de cette simplification est qu'elle permet de calculer facilement les mises à jour de la loi, et l'algorithme EM est ainsi approché (algorithme BEM de Govaert et Nadif). L'inconvénient est qu'un point stationnaire de cet algorithme ne peut être un point stationnaire de la vraisemblance que si le modèle satisfait aux conditions de simplification de l'approximation variationnelle (Gunawardana et Byrne (2005)). Ce qui est parfois réalisé dans certaines conditions asymptotiques, mais qui ne l'est en général pas à distance finie.

# 4 L'algorithme SEM-Gibbs

Nous proposons d'utiliser une version stochastique SEM (Celeux, Chauveau et Diebolt (1996)) de l'EM, dans laquelle l'étape d'estimation est remplacée par la génération d'un échantillon des données manquantes $(\mathbf{w}^{(c)}, \mathbf{z}^{(c)})$ sous la loi des données manquantes conditionnellement aux observations et à l'état en cours $\theta^{(c)}$ du paramètre : on obtient ainsi un pseudo-échantillon complet. L'étape de maximisation recherche le paramètre maximisant la vraisemblance complétée, dans laquelle les variables manquantes sont remplacées par leur tirage.
Pour contourner l'impossibilité du calcul de la loi $p(\mathbf{w}, \mathbf{z}|\mathbf{x}; \theta^{(c)})$, nous proposons de la simuler par un échantillonneur de Gibbs : simulation itérative de $\mathbf{W}$ suivant $p(\mathbf{w}|\mathbf{x}, \mathbf{z}; \theta^{(c)})$, puis de $\mathbf{Z}$ suivant $p(\mathbf{z}|\mathbf{x}, \mathbf{w}; \theta^{(c)})$. En effet, les lois $p(\mathbf{w}|\mathbf{x}, \mathbf{z}; \theta^{(c)})$ et $p(\mathbf{z}|\mathbf{x}, \mathbf{w}; \theta^{(c)})$ sont facilement calculables, et on obtient :

$$p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}) = \prod_i p(z_i|x_{i\cdot}, \mathbf{w}^{(c)}), \quad p(z_i = k|x_{i\cdot}, \mathbf{w}^{(c)}) = \frac{\pi_k \psi_k(x_{i\cdot}, \alpha_{k\cdot})}{\sum_{k'} \pi_{k'} \psi_{k'}(x_{i\cdot}, \alpha_{k'\cdot})}, k = 1, \dots g$$

où $x_{i\cdot}$ désigne la $i^{\text{e}}$ ligne de la matrice $\mathbf{x}$, $\alpha_{k\cdot} = (\alpha_{k1}, \dots, \alpha_{km})$ et

$$\psi_k(x_{i\cdot}, \alpha_{k\cdot}) = \prod_\ell \alpha_{k\ell}^{u_{i\ell}} (1 - \alpha_{k\ell})^{d_\ell - u_{i\ell}}, \quad u_{i\ell} = \sum_j w_{j\ell}^{(c)} x_{ij}, \quad d_\ell = \sum_j w_{j\ell}^{(c)}.$$

De façon similaire :

$$p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c)}) = \prod_j p(w_j|x_{\cdot j}, \mathbf{z}^{(c)}), \quad p(w_j = \ell|x_{\cdot j}, \mathbf{z}^{(c)}) = \frac{\rho_\ell \phi_\ell(x_{\cdot j}, \alpha_{\cdot\ell})}{\sum_{\ell'} \rho_{\ell'} \phi_{\ell'}(x_{\cdot j}, \alpha_{\cdot\ell'})}, \ell = 1, \dots, m$$

où $x_{\cdot j}$ désigne la $j^{\text{e}}$ colonne de la matrice $\mathbf{x}$ , $\alpha_{\cdot\ell} = (\alpha_{1\ell}, \dots, \alpha_{g\ell})$ et

$$\phi_\ell(x_{\cdot j}, \alpha_{\cdot\ell}) = \prod_k \alpha_{k\ell}^{v_{kj}} (1 - \alpha_{k\ell})^{n_k - v_{kj}}, \quad v_{kj} = \sum_i z_{ik}^{(c)} x_{ij}, \quad n_k = \sum_i z_{ik}^{(c)}.$$

On retrouve la définition de $s_{ik}$ (réciproquement $t_{j\ell}$) de Govaert et Nadif dans l'expression de $p(z_i = k|x_{i\cdot}, \mathbf{w}^{(c)})$ (resp. $p(w_j = \ell|x_{\cdot j}, \mathbf{z}^{(c)})$), mais l'espérance conditionnelle $t_{j\ell}$ (resp. $s_{ik}$) dans le calcul de $u_{i\ell}$ (resp. $v_{kj}$) y a été remplacée par le tirage $w_j$ (resp. $z_i$). Les autres paramètres se calculent simplement :

$$\pi_k^{(c+1)} = \frac{\sum_i z_{ik}^{(c)}}{n}, \rho_\ell^{(c+1)} = \frac{\sum_j w_{j\ell}^{(c)})}{d}, \alpha_{k\ell}^{(c+1)} = \frac{\sum_{ij} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} x_{ij}}{\sum_{ij} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)}}.$$

L'algorithme SEM-Gibbs est ainsi défini :

**SEM-Gibbs** : Itération successive de deux étapes, la première étant elle-même une itération d'un schéma de Gibbs pour simuler les données manquantes :

1. Étape E-S : on répète les deux étapes suivantes
   (a) estimation puis tirage de $\mathbf{Z}^{(t+1)}$ suivant la loi $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(t)}; \theta^{(c)})$
   (b) estimation puis tirage de $\mathbf{W}^{(t+1)}$ suivant la loi $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(t+1)}; \theta^{(c)})$
   (c) d'où $w^{(c+1)}$ et $z^{(c+1)}$
2. Étape M : mise à jour de $\theta^{(c+1)}$

Deux variantes de l'algorithme sont également envisagées :

**Variante 1** : on n'attend pas que le schéma de Gibbs atteigne la loi stationnaire : l'étape E-S de l'algorithme précédent n'a qu'un pas. D'où l'itération successive des trois étapes suivantes :

1. Étape E-S-a : estimation puis tirage de $\mathbf{Z}^{(c+1)}$ suivant la loi $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}; \theta^{(c)})$
2. Étape E-S-b : estimation puis tirage de $\mathbf{W}^{(c+1)}$ suivant la loi $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c+1)}; \theta^{(c)})$
3. Étape M : mise à jour de $\theta^{(c+1)}$

**Variante 2** : le paramètre $\theta$ est mis à jour après chaque tirage d'une loi conditionnelle.

1. Étape E-S-a : estimation puis tirage de $\mathbf{Z}^{(c+1)}$ suivant la loi $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}; \theta^{(c)})$
2. Étape M : mise à jour de $\tilde{\theta}^{(c)}$
3. Étape E-S-b : estimation puis tirage de $\mathbf{W}^{(c+1)}$ suivant la loi $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c+1)}; \tilde{\theta}^{(c)})$
4. Étape M : mise à jour de $\theta^{(c+1)}$

# 5 Conclusion

Dans cet exposé, nous comparerons les résultats entre les trois variantes de l'algorithme SEM présentées ci-dessus et l'algorithme BEM, aussi bien au niveau de la qualité des résultats, qu'au niveau de la rapidité d'exécution.

## Bibliographie

[1] Govaert, G. et Nadif M. (2008) Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52, 3233–3245.

[2] Dempster, A.P. ; Laird, N.M. et Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1), 1–38.

[3] Keribin, C. (2009) Les méthodes bayésiennes variationnelles et leur application en neuroimagerie : une étude de l'existant. *Rapport de Recherche INRIA 7091.*

[4] Gunawardana, A. et Byrne W. (2005) Convergence theorems for Generalized Alternating Minimization procedures. *Journal of Machine learning Research*, 6, 2049–2073.

[5] Celeux, G. ; Chauveau D. et Diebolt J. (1996) Stochastic versions of the EM algorithm. *Journal of Statist. Comput. Simulation* 55, 287–314